ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 8, No. 2, 2013, *pp*.119-129

Privacy Preserving Three-Layer Naïve Bayes Classifier for Vertically Partitioned Databases

Alka Gangrade [1] and Ravindra Patel [2]

[1] Technocrats Institute of Technology, Bhopal, India, Email: alkagangrade@yahoo.co.in
[2] Dept. of M.C.A., U.I.T., R.G.P.V., Bhopal, India, Email: ravindra@rgtu.net

(*Received October 28, 2012, accepted February 29, 2013)*

Abstract. Data mining is the extraction of the hidden information from large databases. It is a powerful technology to explore important information in the data warehouse. Privacy preservation is a significant problem in the field of data mining. It is more challenging when data is distributed among different parties. In this paper, we address the problem of privacy preserving three-layer Naïve Bayes classification over vertically partitioned data. Our approach is based on Secure Multiparty Computation (SMC). We use secure multiplication protocol to classify the new tuples. In our protocol, secure multiplication protocol allows to meet privacy constraints and achieve acceptable performance and our classification system is very efficient in term of computation and communication cost.

Keywords: Privacy preserving, Naïve Bayes classification, probability, secure multiplication protocol.

1. Introduction

Boundary heat

Classification is a popular data mining technique used to predict group membership for data tuples. In classification rule mining, a set of database tuples act as a training sample and it is analyzed to produce a model of the data or classifier that can be used for classifying a new tuple. The popular classification rule mining techniques are decision trees, neural networks, Naïve Bayesian classifiers etc. Preserving privacy against data mining algorithms is a new research area. Privacy preserving data mining is the emerging field that protects sensitive data. The goal of privacy preserving classification is to build precise classifiers without disclosing personal information in the data being mined.

1.1. Naïve Bayesian Classification

Bayesian classification is based on Bayes' theorem [1]. A simple Bayesian classifier is known as the Naïve Bayesian classifier, to be comparable in performance with decision tree and selected neural network classifier. Bayesian classifiers have also exhibited high accuracy and speed when applied to large database.

Bayes' theorem is

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \tag{1}$$

Where H is some hypothesis, such as that the data tuple X belongs to a specified class 'C'. For classification problems, we want to determine P (H|X), the probability that the hypothesis H holds given the "evidence" or observed data tuple X.

P (H|X) is the posterior probability of H conditioned on X.

P (H) is the prior probability of H. For our example; this is the probability that any given customer will buy a computer, regardless of age, income or any other information.

P (X|H) is the posterior probability of X conditioned on H.

Naïve Bayes is extremely effective but straightforward classifier. Due to this combination of straightforward and effectiveness it is used as a baseline standard by which other classifiers are measured. Naïve Bayes represents each class with a probabilistic summary, and classify each new tuple with the most likely class. It provides a flexible way for dealing with any number of attributes or classes, and is based on probability theory. It is fast learning algorithm that examines all its training input. It has been established to achieve unexpectedly well in a wide variety of problems despite of the simple nature of the model. With various enhancements it is highly effective, and receives practical use in many applications for example content based filtering and text categorization.

For preserving privacy we use the framework defined in Secure Multiparty Computation [2], and several primitives from the Secure Multiparty Computation contents. Complete details of Naïve Bayes classification algorithms can be found in [3]. We assume that the basic formulae are well known. In order to construct a privacy preserving Naïve Bayesian classifier, we must concentrate on two issues, how to calculate the probability or model parameter for each attribute and how to classify a new tuple [4, 5, 6]. The following subsections provide details on both issues. The protocol presented below is quite efficient.

1.2. Our Contributions

Our main contributions in this paper are as follows:

- We present a novel privacy preserving Naïve Bayes classifier for vertically partitioned databases.
- It classifies new tuple by using secure multiplication protocol. We propose a new protocol.

1.3. Organization of the paper

The rest of the paper is organized as follows. In Section 2, we discuss the related work. Section 3, describes proposed work of our novel privacy preserving Naïve Bayes classification model for vertically partitioned data. Section 3.1 describes architecture of our model and secure multiplication protocol. Section 3.2 sets some assumptions. Section 3.3 describes formal algorithms of our proposed work. In Section 4, we present our calculation and results that are conducted by using our proposed model on real-world data sets. In Section 5, we conclude our paper with the discussion of the future work.

2. Related Work

Privacy preserving data mining has been an active research area for a decade. A lot of work is going on by the researcher on privacy preserving classification in distributed data mining. The first Secure Multiparty Computation (SMC) problem was described by Yao [7]. SMC allows parties with similar background to compute result upon their private data, minimizing the threat of disclosure was explained [8].

There have been several approaches to support privacy preserving data mining over multi-party without using third parties [9, 10]. Some techniques, review and evaluation of privacy preserving algorithms also presented in [9]. Various tools discussed and how they can be used to solve several privacy preserving data mining problems [11]. We now give some of the related work in this area. Previous work in privacy preserving data mining has addressed some issues. The aim is to preserve customer privacy by distorting the data values presented in [10]. D. Agrawal and C. C. Aggarwal designed various algorithms for improving this approach [12].

Classification is one of the most widespread data mining problems come across in real life. General classification techniques have been extensively studied for over twenty years. The classifier is usually represented by classification rules, decision trees, Naïve Bayes classification and neural networks. First ID3 decision tree classification algorithm is proposed by Quinlan [13]. Lindell and Pinkas proposed a secure algorithm to build a decision tree using ID3 over horizontally partitioned data between two parties using SMC [14]. A novel privacy preserving distributed decision tree learning algorithm [15] that is based on Shamir [16]. The ID3 algorithm is scalable in terms of computation and communication cost, and therefore it can be run even when there is a large number of parties involved and eliminate the need for third party and propose a new method without using third parties. A generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data distributed over two or more parties introduced in [17, 18, 19, 20] and horizontally partitioned data distributed over multi parties introduced in [21, 22]. Privacy preserving Naïve Bayes classification for horizontally partitioned data introduced in [4] and vertically partitioned data introduced in [5, 6]. Centralized Naïve Bayes classification probability calculation is introduced in [23].

3. Proposed Work

This paper addresses classification over vertically partitioned data, where different parties hold different attributes. We consider the case where all party holds the class attributes. In this case, all party calculates